

Round-off error propagation in the integration of ordinary differential equations by one-step methods

MÁTYÁS ARATÓ

Dedicated to Professor Béla Szökefalvi-Nagy on his 70-th birthday

The connection between round-off errors in the integration of a system of ordinary differential equations by one-step methods and stochastic differential equations with respect to a wide sense Wiener process is examined. The generalization of RADEMACHER's [8] and HENRICI's theorems [5] is given using Ito's integral. Under some weak conditions on the behavior of local round-off errors one can calculate the mean value and variance of the propagated round-off error. It turns out that to any system of differential equations a stochastic system of equations is related which describes the round-off error propagation.

The distribution of the propagated round-off error, \mathbf{r}_x , depends on the distribution of local errors; the conditions of Gaussiennes are also given. This is a sharpening of a special result of Henrici. We take advantage of the optimal filtering equations to give the expected value and variance of round-off error. This problem was studied earlier for the best approximate solutions of linear algebraic systems (see TITONOV [9], [10], LIPTSER and SHIRYAEV [6]). The natural question on the distribution of $\max \|\mathbf{r}_x\|$, in $0 \leq x \leq b$, is also examined and using some recent results of NOVIKOV [7] we answer it in the one-dimensional case. The description of stochastic equations in multistep methods and especially in predictor-corrector methods remains open.

1. Introduction. Let us consider the following first order vector initial value problem

$$(1) \quad \mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)), \quad \mathbf{y}(x_0) = \mathbf{y}(0), \quad 0 < x < b,$$

where \mathbf{y} and \mathbf{f} are column vectors

$$\mathbf{y} = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^k \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} f^1 \\ f^2 \\ \vdots \\ f^k \end{pmatrix}.$$

(The superscripts always denote indices and asterisk indicates the transposition of a matrix or vector; e.g. y^* means a row vector $y^* = (y^1, \dots, y^k)$.) If v is a vector with real or complex components the norm is given by $\|v\| = |v^1| + \dots + |v^k|$.

A one-step method for the solution of the initial value problem is defined by the formulas

$$(2) \quad y_0 = y(0), \quad y_{n+1} = y_n + h\Phi(x, y_n; h), \quad h > 0,$$

where $\Phi(x, y; h)$ is called the increment function and is chosen so as to approximate $(y(x+h) - y(x))/h$ as well as possible. We assume that $\Phi(x, y; h)$ is continuous and that there exists a constant L_1 such that

$$(3) \quad \|\Phi(x, \tilde{y}; h) - \Phi(x, y; h)\| \leq L_1 \|\tilde{y} - y\|,$$

for all points $(x, \tilde{y}; h)$ and $(x, y; h)$, $h < h_0$ (h_0 is fixed).

The discretization error e_n is defined as

$$(4) \quad e_n = y_n - y(x_n)$$

where $y(x_n)$ denotes the solution of the initial value problem at x_n .

The round-off error r_n is defined as the difference between y_n and its numerical approximation \hat{y}_n , i.e.,

$$r_n = \hat{y}_n - y_n,$$

and it depends on the local errors and the kind of arithmetic used in computer. r_n fulfils a stochastic difference equation, depending on $\Phi(x, y; h)$ (see (12)). Under some weak conditions on Φ the continuous function r_x of x is a solution of the stochastic differential equation (19) (see below), which we call the related stochastic equation to (1). The investigation of equation (19) is the main goal of this paper as in earlier papers only approximations and estimations were given (see HENRICI [5]).

As a new and natural aspect we exercise the distribution of $\max_{a \leq x \leq b} \|r_x\|$, which is an effective measure of the error behavior. Theorem 4 gives the answer in the one dimensional case, with sharp bounds instead of the estimation of the mean, as it is used in the literature.

2. Derivation of the related stochastic equation. Using Euler's method, i.e., $\Phi(x, y) = f(x, y)$, it is easy to prove that

$$(5) \quad e_{n+1} = e_n + h[f(x_n, y(x_n)) - f(x_n, y_n)] + \frac{h^2}{2} y''(\xi),$$

$$\|e_{n+1}\| \leq (1 + hL)\|e_n\| + \frac{h^2}{2} k,$$

and this gives, assuming $\|y''\| < K$ and $\|f(x, y) - f(x, \tilde{y})\| < L\|y - \tilde{y}\|$, the following estimation (see e.g. HENRICI [5])

$$(6) \quad \|e_n\| \leq \frac{hK}{L} [e^{L(x_n - x_0)} - 1].$$

Let \hat{y}_n denote the numerical approximation of y_n . The local error ε_n at step n is induced by computer round-off (or chopping) and the inherent error by inaccuracy of evaluation of function $\Phi(x_n, y_n; h)$.

Instead of equation (2) we have

$$(7) \quad \hat{y}_{n+1} = \hat{y}_n + h\Phi(x_n, \hat{y}_n; h) + \tilde{\varepsilon}_{n+1},$$

and the accumulated round-off error $r_n = \hat{y}_n - y_n$ fulfils the equation, subtracting (7) and (2)

$$(8) \quad r_{n+1} = r_n + h[\Phi(x_n, \hat{y}_n; h) - \Phi(x_n, y_n; h)] + \tilde{\varepsilon}_{n+1}.$$

This means that the accumulated round-off error is not simply the sum of local round-off errors. It depends on the kind of arithmetic used in computer, the way in which the machine rounds, the order in which the arithmetic operations are performed and on the numerical procedures being used. As over an extended interval the loss of accuracy may be serious, it is desirable to obtain estimates by making some statistical assumptions on the behaviour of local round-off errors $\tilde{\varepsilon}_n$.

It is known that by double precision the possible gain in accuracy can be very significant, but we have a loss in performance and efficiency.

A crude bound for the accumulated round-off error r_n can be obtained from (8) if we assume that

$$(9) \quad \|\tilde{\varepsilon}_n\| \leq \varepsilon, \quad n = 1, 2, \dots;$$

namely, using (3) we get

$$(10) \quad \|r_n\| \leq \frac{\varepsilon}{hL_1} [e^{L_1(x_n - x_0)} - 1].$$

Comparing (10) and (6) we see, as the accuracy of numerical integration depends upon the discretization error and the accumulated rounding error, that it is impossible to keep both of the errors small. To keep the discretization error small, we will normally choose the stepsize h small. On the other hand, the smaller h is taken, the more integration steps we shall have to perform, and the greater the rounding error is likely to be. An optimum value of the stepsize h must exist but it seems difficult to find it in practice.

In order to obtain realistic statements concerning the behaviour of the propagated round-off errors, from here we shall assume that the local round-off errors $\tilde{\varepsilon}_n$ are random variables. In the simplest case $\tilde{\varepsilon}_n$ is a white noise process, i.e., $\text{cov}(\tilde{\varepsilon}_n, \tilde{\varepsilon}_m) = 0$, if $n \neq m$.

Let us assume further that $\mu_n = E\tilde{\epsilon}_n$ for which

$$(11) \quad \begin{pmatrix} \mu_n^1 \\ \vdots \\ \mu_n^k \end{pmatrix} = \mu(h) \mathbf{p}(x_n),$$

where $\mu(h)/h \rightarrow \mu$, $h \downarrow 0$, and μ is a constant and $\mathbf{p}(x)$ is a known vector function with components which are smooth functions of x . Let

$$(12) \quad \text{cov}(\tilde{\epsilon}_n, \tilde{\epsilon}_m) = E(\tilde{\epsilon}_n - \mu_n)(\tilde{\epsilon}_m - \mu_m)^* = \begin{cases} B_z(x_n)h & \text{if } n = m, \\ 0 & \text{if } n \neq m. \end{cases}$$

And assuming the smoothness of Φ let

$$(13) \quad \Phi(x_n, \hat{y}_n; h) - \Phi(x_n, y_n; h) = G(x_n) \mathbf{r}_n + \theta_n h,$$

where the matrix $G(x_n)$ can be expressed by the derivatives of Φ (see [5]). Then (8) can be rewritten in the form

$$(14) \quad \mathbf{r}_{n+1} = (I + hG(x_n))\mathbf{r}_n + \epsilon_{n+1}$$

where $\epsilon_n = \tilde{\epsilon}_n + \theta_n h$, or

$$(14') \quad \mathbf{r}_{n+1} - \mathbf{r}_n = hG(x_n)\mathbf{r}_n + \mu(h)\mathbf{p}(x_{n+1}) + B_z^{1/2}(x_{n+1})\epsilon_{n+1}^0,$$

where $\epsilon_{n+1} - \mu(h)\mathbf{p}(x_{n+1}) = B_z^{1/2}(x_{n+1})\epsilon_{n+1}^0$, with $\text{cov}(\epsilon_n^0, \epsilon_n^0) = hI$. Now approximating the process ϵ_{n+1}^0 by a wide sense Wiener process (see the definition in LIPTSER—SIRYAEV [6] Section 15) we get that $\mathbf{r}_n = \mathbf{r}(x_n)$ has the stochastic differential

$$(15) \quad d\mathbf{r}_x = G(x)\mathbf{r}_x dx + \mu\mathbf{p}(x) dx + B_w^{1/2}(x) d\mathbf{w}(x), \quad \mathbf{r}_0 = 0,$$

where $\mathbf{w}(0) = 0$, $E\mathbf{w}(x) = 0$, $E\mathbf{w}(x_1)\mathbf{w}^*(x_2) = I \min(x_1, x_2)$. It is clear that any Wiener process is a wide sense Wiener process at the same time.

Equation (15) can be considered as the linear equation

$$(16) \quad \mathbf{r}_x = \int_0^x [\mu\mathbf{p}(u) + G(u)\mathbf{r}_u] du + \int_0^x B_w^{1/2}(u) d\mathbf{w}(u), \quad \mathbf{r}_0 = 0,$$

with the unique continuous solution (in mean square)

$$(17) \quad \mathbf{r}_x = \Phi_0(x) \left\{ \int_0^x (\Phi_0(s))^{-1} \mu\mathbf{p}(s) ds + \int_0^x (\Phi_0(s))^{-1} B_w^{1/2}(s) d\mathbf{w}(s) \right\}, \quad \mathbf{r}_0 = 0,$$

where $\Phi_0(x)$ is the fundamental matrix

$$(18) \quad \frac{d\Phi_0(x)}{dx} = G(x)\Phi_0(x), \quad \Phi_0(0) = I_{k \times k},$$

i.e.,

$$(18') \quad \Phi_0(x) = \exp \left\{ \int_0^x G(u) du \right\}.$$

From the smoothness assumption on $p(x)$ and $G(x)$ it follows that

$$(19) \quad \int_0^b |p^i(x)| dx < \infty, \quad \int_0^b |g_{ij}(x)| dx < \infty, \quad \int_0^b b_{ij}(x) dx < \infty, \\ i, j = 1, 2, \dots, k.$$

The following statement immediately follows from Theorem 15.1 in [6], where

$$(20) \quad E\mathbf{r}_x = \mathbf{m}(x), \quad B(x, u) = E(\mathbf{r}_x - \mathbf{m}(x))(\mathbf{r}_u - \mathbf{m}(u))^*.$$

Theorem 1 (see HENRICI [5]). *Suppose that the conditions (19) hold and \mathbf{r}_x fulfils the stochastic differential equation (15) with a wide sense Wiener process $\mathbf{w}(x)$. Then the vector $\mathbf{m}(x)$ and the matrix $B(x)$ are solutions of the differential equations*

$$(21) \quad \frac{d\mathbf{m}(x)}{dx} = \mu p(x) + G(x)\mathbf{m}(x),$$

$$(22) \quad \frac{dB(x)}{dx} = G(x)B(x) + B(x)G^*(x) + B_w(x).$$

The matrix $B(x, u)$ is given by the formula

$$(23) \quad B(x, u) = \begin{cases} \Phi(u, x)B(u), & u \leq x, \\ B(x)(\Phi(x, u))^*, & u \geq x, \end{cases}$$

and

$$\Phi(u, x) = \Phi_0(x)(\Phi_0(u))^{-1}, \quad u \leq x.$$

Proof. Taking expectations of both sides in (16) we get (21) and from (17) it follows that

$$(24) \quad \mathbf{m}(x) = \Phi_0(x) \left\{ \int_0^x (\Phi_0(u))^{-1} \mu p(u) du \right\}, \quad \mathbf{m}(0) = \mathbf{0}.$$

Let $\mathbf{r}_x - \mathbf{m}(x) = \tilde{\mathbf{r}}_x$, then from (17) and (24) one can get

$$(25) \quad \tilde{\mathbf{r}}_x = \Phi_0(x) \left\{ \int_0^x (\Phi_0(u))^{-1} B_w^{1/2}(u) d\mathbf{w}(u) \right\},$$

and

$$(26) \quad E(\tilde{\mathbf{r}}_x \tilde{\mathbf{r}}_x^*) = \Phi_0(x) E \left\{ \int_0^x (\Phi_0(u))^{-1} B_w^{1/2}(u) d\mathbf{w}(u) \cdot \left[\int_0^x (\Phi_0(u))^{-1} B_w^{1/2}(u) d\mathbf{w}(u) \right]^* \right\} \Phi_0^*(x) = \Phi_0(x) \int_0^x (\Phi_0(u))^{-1} B_w(u) (\Phi_0^{-1}(u))^* du \Phi_0^*(u).$$

By differentiating and taking into account (18) one can get (22).

To establish (23) let $x \cong u$. Then

$$(27) \quad \begin{aligned} E\tilde{\mathbf{r}}_x \tilde{\mathbf{r}}_n^* &= \Phi_0(x) E \left(\int_0^x \Phi_0^{-1}(u) B_w^{1/2}(u) d\mathbf{w}(u) \left[\int_0^x \chi(u \cong s) \Phi_0^{-1}(s) B_w^{1/2}(s) d\mathbf{w}(s) \right] \right)^* \\ &\cdot (\Phi_0(u))^* = \Phi_0(x) \Phi_0(u) \left\{ \int_0^u \Phi_0^{-1}(s) B_w(s) (\Phi_0^{-1}(s))^* ds \right\} \Phi_0^*(u) = \Phi(u, x) B(u), \end{aligned}$$

which proves the theorem.

The following reverse statement is also true.

Theorem 2. Let $\mathbf{r}_x = (r_x^1, \dots, r_x^k)$, $0 \leq x \leq b$, be a random vector process with given first two moments

$$(28) \quad \mathbf{m}(x) = E\mathbf{r}_x, \quad B(x, u) = E(\mathbf{r}_x - \mathbf{m}(x))(\mathbf{r}_u - \mathbf{m}(u))^*.$$

Assume that $B_w(x)$ is nonnegative definite and the following assumptions are satisfied:

a) The elements of the vector $\mathbf{p}(x)$ and the matrices $B(x)$, $B_w(x)$ are Lebesgue integrable.

b) The matrix $B(x) = B(x, x)$ has continuous elements and

$$(29) \quad B(x) = \int_0^x [G(u)B(u) + B(u)G^*(u)] du + \int_0^x B_w(u) du, \quad B(0) = 0,$$

c) $\mathbf{m}(x)$ has continuous components and

$$(30) \quad \mathbf{m}(x) = \int_0^x [\mu \mathbf{p}(u) + G(u)\mathbf{m}(u)] du.$$

Then there exists a wide sense Wiener process $\tilde{\mathbf{w}}^*(x) = (\tilde{w}^1(x), \dots, \tilde{w}^k(x))$ such that for all x , $0 \leq x \leq b$,

$$(31) \quad \mathbf{r}_x = \int_0^x [\mu \mathbf{p}(u) + G(u)\mathbf{r}_u] du + \int_0^x B_w^{1/2}(u) d\tilde{\mathbf{w}}(u)$$

The proof immediately follows from Theorem 15.2 in [6].

3. The distribution of the maximum of round-off error. The interpretation of the results in Section 2 is the following. If $E\varepsilon_n = \mu_n = \mu h \mathbf{p}(x_n)$, then we have

$$(32) \quad E\mathbf{r}_n = (\mathbf{m}(x_n) + O(h)),$$

where $\mathbf{m}(x)$ is the solution of the initial value problem

$$(33) \quad \mathbf{m}'(x) = G(x)\mathbf{m}(x) + \mu \mathbf{p}(x),$$

with the assumption that the matrix $G(x)$ is given by

$$(34) \quad \Phi(x_n, \mathbf{y}_n; h) - \Phi(x_n, \tilde{\mathbf{y}}_n; h) = G(x_n)(\mathbf{y}_n - \tilde{\mathbf{y}}_n) + \varepsilon \Theta_n, \quad \varepsilon > 0, \quad \|\Theta_n\| < 1.$$

The process r_x is stationary if $G(x)=A$ and in this case $B'_x=0$ and $B_x=B_0$ is the solution of the equation (see [1], [2])

$$(35) \quad AB_0 + B_0A = -B_w,$$

i.e., r_x has a normal distribution with parameters $(0, B_0)$. Note that if h is small, $B_\varepsilon \cong B_w h$ and from (35) we see that

$$B_0 \sim \frac{1}{h} (B_\varepsilon + O(h)), \quad (\text{see (10)}).$$

In many cases we are interested in the behavior of the round-off error on the whole interval, i.e., in the values

$$P\left\{\sup_{0 \leq x \leq b} \|r_x\| \leq \varepsilon\right\}, \quad P\{\|r_x\| \leq g(x), \quad 0 \leq x \leq b\},$$

which gives a better estimation than (32). For simplicity let us consider the one dimensional case and we assume $p(x)=0$. Let $G(x)$ be given by

$$(36) \quad G(x) = \frac{m'(x)}{m(x)},$$

where $m(x)$ is a positive continuous function for $x \geq 0$.

We prove the following statements.

Lemma. *Let $w(x)$ be the standard Wiener process, $w(0)=0$, $EW(x)=0$, $EW^2(x)=x$, and let $m(x)$ be a positive continuous function, and let $G(x)$ be defined by (36). Then*

$$(37) \quad r_x = m(x) \int_0^x m^{-1}(u) B_w^{1/2}(u) dw(u),$$

where

$$(16') \quad dr_x = G(x)r_x dx + B_w^{1/2}(x) dw(x).$$

Proof. By Ito's formula it is easy to get from (37) that

$$(38) \quad dr_x = \frac{m'(x)}{m(x)} r_x dx + B_w^{1/2}(x) dw(x),$$

and comparing with (16') one can get the statement.

Theorem 3. *Let r_x be the process defined by (16), where $p(x)=0$, and further let $m(x)$ be a positive function with continuous $m''(x)$ ($x \geq 0$). Then for all $0 \leq b < \infty$*

$$(39) \quad \frac{8}{3\pi} \leq P\{|r_x| \leq km(x), \quad 0 \leq x \leq b\} \exp \left\{ \frac{\pi^2}{8k^2} \int_0^b m^{-2}(x) B_w(x) dx \right\} \leq \frac{4}{\pi}.$$

Proof. From the Lemma it follows that

$$\begin{aligned}
 P\{|r_x| \leq km(x), 0 \leq x \leq b\} &= P\left\{\left|\int_0^x m^{-1}(u) B_w^{1/2}(u) dw(u)\right| \leq k, 0 \leq x \leq b\right\} = \\
 (40) \quad &= P\{|\tilde{w}(u)| \leq k, 0 \leq u \leq \int_0^b m^{-2}(u) B_w(u) du\},
 \end{aligned}$$

where $\tilde{w}(u)$ is a new Wiener process obtained by the "time" change

$$(41) \quad u = \int_0^x m^{-2}(s) B_w(s) ds$$

from the stochastic integral $\int_0^x m^{-1}(s) B_w^{1/2}(s) dw(s)$, (see [4]). But for the Wiener process the following representation is well known ([3], p. 330)

$$(42) \quad P\{|w(u)| \leq k, 0 \leq u \leq c\} = \frac{4}{\pi} \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} \exp\left[-(2n+1)^2 \frac{\pi^2 c}{8k^2}\right].$$

In (42) on the right side there is an alternating series and we have the following estimates

$$\begin{aligned}
 (43) \quad \frac{4}{\pi} \left[\exp\left(-\frac{\pi^2}{8k^2} c\right) - \frac{1}{3} \exp\left(-\frac{9\pi^2}{8k^2} c\right) \right] &\leq P\{|w(u)| \leq k, 0 \leq u \leq c\} \leq \\
 &\leq \frac{4}{\pi} \exp\left(-\frac{\pi^2}{8k^2} c\right),
 \end{aligned}$$

which together with (40) gives (39), and this proves the statement.

Remark 1. In the case $m(x) = ax + \tilde{b}$, $a > 0$, $\tilde{b} \geq 0$, and $B_w = \sigma^2$ we have

$$\begin{aligned}
 (44) \quad \frac{8}{3\pi} \exp\left(-\frac{\pi^2}{8k^2} \cdot \frac{\sigma^2}{a} \left[\frac{1}{\tilde{b}} - \frac{1}{ab + \tilde{b}}\right]\right) &\leq P\{|r_x| \leq km(x), 0 \leq x \leq b\} \leq \\
 &\leq \frac{4}{\pi} \exp\left(-\frac{\pi^2}{8k^2} \cdot \frac{\sigma^2}{a} \left[\frac{1}{\tilde{b}} - \frac{1}{ab + \tilde{b}}\right]\right).
 \end{aligned}$$

Remark 2. Let $m(x) = a(x+1)^{1/2}$, $a > 0$ and $B_w = \sigma^2$, then

$$\begin{aligned}
 (45) \quad \frac{8}{3\pi} (1+b)^{-\frac{\pi^2}{8} \frac{\sigma^2}{ka^2}} &\leq P\{|r_x| \leq ka(1+x)^{1/2}, 0 \leq x \leq b\} \leq \\
 &\leq \frac{4}{\pi} (1+b)^{-\frac{\pi^2}{8} \frac{\sigma^2}{ka^2}}.
 \end{aligned}$$

This formula gives the following asymptotic result for the stopping time

$$\tau = \inf \{x: |r_x| \geq km(x)\},$$

$$P(\tau > b) = P\{|r_x| \leq km(x), 0 \leq x \leq b\} \sim c_0(1+b)^{-\frac{c_1}{a^2}}, \quad c_0, c_1 > 0,$$

where $a \sim 0$.

Remark 3. Estimates for the probability that the process r_x will not exit to a one-sided moving boundary can be handled in the same way (see NOVIKOV [7]).

References

- [1] M. ARATÓ, On the statistical examination of continuous state Markov processes. II, *Magyar Tud. Akad. Mat. Fiz. Oszt. Közl.*, **14** (1964), 137—159 (Hungarian); English transl.: *Selected Transl. Math. Stat. and Probab.*, Amer. Math. Soc., **14** (1978), 227—251.
- [2] M. ARATÓ, *Linear stochastic systems with constant coefficients*, Lecture Notes in Control and Information, vol. 45, Springer-Verlag (Berlin, 1982).
- [3] W. FELLER, *An introduction to probability theory and its applications*, Vol. 2, Wiley (New York, 1966).
- [4] J. GIHMAN and A. SKOROKHOD, *Stochastic differential equations*, Ergebnisse der Mathematik und ihrer Grenzgebiete, B. 72, Springer-Verlag (Berlin, 1972).
- [5] P. HENRICI, *Discrete Variable Methods in Ordinary Differential Equations*, J. Wiley (New York, 1962).
- [6] R. LIPTSER and A. SHIRYAEV, *Statistics of random processes*, Nauka (Moscow, 1974) (Russian).
- [7] A. NOVIKOV, On estimates and the asymptotic behavior of nonexit probabilities of a Wiener process to a moving boundary, *Mat. Sbornik*, **110** (152) (1979), No. 4, 539—550 (Russian).
- [8] H. RADEMACHER, On the accumulation of errors in processes of integration on high-speed calculating machines, *Annals Comput. Lab. Harvard Univ.*, **16** (1948), 176—187.
- [9] A. N. TIHONOV, On approximate systems of linear algebraic equations, *Computer Math. and Math. Physics*, **20** (1980), 1373—1383 (Russian).
- [10] A. N. TIHONOV, On the normal solutions of approximate systems of linear algebraic equations, *Dokl. Akad. Nauk SSSR*, **254** (1980), No. 3, 549—554 (Russian).